IFW

Attorney Docket No. YOR920000429US1

AF/2121

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**



**Patent Application**

Applicant(s): C.C. Aggarwal et al
Docket No.: YOR920000429US1
Serial No.: 09/686,115
Filing Date: October 11, 2000
Group: 2121
Examiner: Kelvin E. Booker

I hereby certify that this paper is being deposited on this date with the U.S. Postal Service as first class mail addressed to the Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450

Signature: Lisa L. Culpis Date: August 10, 2004

Title: Methods and Apparatus for Outlier Detection for High Dimensional Data Sets

---

TRANSMITTAL OF APPEAL BRIEF

Mail Stop Appeal Brief - Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

Submitted herewith are the following documents relating to the above-identified patent application:

(1) Appeal Brief in triplicate (original and two copies); and

(2) Copy of Notice of Appeal, filed on June 8, 2004, with copy of stamped return postcard indicating receipt of Notice by PTO on June 10, 2004.

Please charge **International Business Machines Corporation Deposit Account No. 50-0510** the amount of $330 to cover this submission under 37 CFR §1.17(c). In the event of non-payment or improper payment of a required fee, the Commissioner is authorized to charge or to credit **Deposit Account No. 50-0510** as required to correct the error. A duplicate copy of this letter and two copies of the Appeal Brief are enclosed.

Respectfully submitted,

Robert W. Griffith

Date: August 10, 2004

Robert W. Griffith
Reg. No. 48,956
Attorney for Applicant(s)
Ryan, Mason & Lewis, LLP
90 Forest Avenue
Locust Valley, NY 11560
(516) 759-4547

Attorney Docket No. YOR920000429US1

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

**Patent Application**



Applicant(s): C.C. Aggarwal et al.
Docket No.: YOR920000429US1
Serial No.: 09/686,115
Filing Date: October 11, 2000
Group: 2121
Examiner: Kelvin E. Booker

Title: Methods and Apparatus for Outlier Detection for High Dimensional Data Sets

---

## APPEAL BRIEF

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313

Sir:

Applicants (hereinafter referred to as "Appellants") hereby appeal the final rejection of claims 1-3, 6-13, 16-23 and 26-30 of the above referenced application.

## REAL PARTY IN INTEREST

The present application is assigned to International Business Machines Corp., as evidenced by an assignment recorded January 4, 2001 in the U.S. Patent and Trademark Office at Reel 11418, Frame 0019. The assignee, International Business Machines Corp., is the real party in interest.

## RELATED APPEALS AND INTERFERENCES

There are no known related appeals and interferences.

## STATUS OF CLAIMS

Claims 1-30 are pending in the present application. Claims 1-3, 6-13, 16-23 and 26-30 stand rejected under 35 U.S.C. §102(a) and claims 4, 5, 14, 15, 24 and 25 are allowable. Claims 1-3, 6-13, 16-23 and 26-30 are appealed.

## STATUS OF AMENDMENTS

There have been no amendments filed subsequent to the final rejection.

## SUMMARY OF INVENTION

The invention relates to outlier detection in high dimensional data and, more particularly, to methods and apparatus for performing such detection in accordance with various high dimensional data domain applications where it is important to be able to find and detect outliers which deviate considerably from the rest of the data (Specification, page 1, lines 4-7).

The present invention provides methods and apparatus for outlier detection which find outliers by observing the density distributions of projections from the data. A point is considered an outlier, if in some lower dimensional projection, it is present in a local region of abnormally low density. Specifically, the invention defines outliers for data by looking at those projections of the data which have abnormally low density (Specification, page 4, lines 18-23).

Accordingly, in an illustrative aspect of the invention, a method of detecting one or more outliers in a data set comprises the following steps. First, one or more sets of dimensions and corresponding ranges (e.g., patterns) in the data set, which are sparse in density (e.g., have an abnormally low presence which cannot be justified by randomness) are determined. Then, one or more data points (e.g., records) in the data set which contain these sets of dimensions and corresponding ranges are determined, the one or more data points being identified as the one or more outliers in the data set (Specification, page 4, line 24 through page 5, line 3).

A diagram showing various patterns of data sets illustrating outlier detection issues are shown in FIG. 1. A flow diagram illustrating an overall process for outlier detection is shown in FIG. 3. More specifically, FIG. 3 describes the steps of first encoding the transactions as strings, and then running the iterative genetic algorithm process on the strings in order to find the appropriate outlier

projections. FIG. 4 is a flow diagram illustrating a process for determining the encoding for each of the records in the database, and FIGs. 5-7 are flow diagrams illustrating procedures for selection, crossover and solution recombination, and mutation, used by a genetic outlier detection algorithm.

Thus, methods and apparatus of the present invention are provided for outlier detection in databases by determining sparse low dimensional projections, which are used for the purpose of determining which points are outliers. The methodologies of the invention are very relevant in providing a novel definition of exceptions or outliers for the high dimensional domain of data (Specification, page 17, lines 4-8).

## ISSUES PRESENTED FOR REVIEW

(I) Whether claims 1-3, 6-9, 11-13, 16-19, 21-23 and 26-29 are properly rejected under 35 U.S.C. §102(a) as being anticipated by Knorr et al., "Distance-Based Outliers: Algorithms and Applications" (hereinafter "Knorr").

(II) Whether claims 10, 20 and 30 are properly rejected under 35 U.S.C. §102(a) as being anticipated by Sheikholeslami et al., "WaveCluster: A Wavelet-Based Clustering Approach for Spatial Data in Very Large Databases" (hereinafter "Sheikholeslami").

## GROUPING OF CLAIMS

Claims 1-3, 6-13, 16-23 and 26-30 do not stand or fall together. More particularly, claims 1-3, 11-13 and 21-23 stand or fall together, claims 6-9, 16-19 and 26-29 stand or fall together, and claims 10, 20 and 30 stand or fall together.

## ARGUMENT

Appellants incorporate by reference herein the disclosure of all previous responses filed in the present application, namely, responses dated December 30, 2003 and June 8, 2004. Sections (I) and (II) to follow will respectively address issues (I) and (II) presented above.

(I) With regard to the rejection of claims 1-3, 6-9, 11-13, 16-19, 21-23 and 26-29 under 35 U.S.C. §102(a) as being anticipated by Knorr, Appellants assert that Knorr fails to teach or suggest each and every element respectively recited in claims 1-3, 6-9, 11-13, 16-19, 21-23 and 26-29.

It is well-established law that "[a] claim is anticipated only if each and every element as set forth in the claims is found, either expressly or inherently described, in a single prior art reference." See, e.g., *Verdegaal Bros. v. Union Oil Co. of California*, 814 F.2d 628, 631, 2 U.S.P.Q.2d 1051, 1053 ( Fed. Cir. 1987). See also, M.P.E.P. §2131. Appellants assert that the §102(a) rejection of claims 1-3, 6-9, 11-13, 16-19, 21-23 and 26-29 based on Knorr clearly fails to meet the above legal requirements for anticipation. Support for this assertion follows.

Independent claims 1, 11 and 21 recite techniques for detecting one or more outliers in a data set. One or more sets of dimensions and corresponding ranges in the data set, which are sparse in density, are determined. One or more data points in these sets are identified as the outliers in the data set.

The present invention identifies outliers by observing the density distributions of projections from the data. It considers a point to be an outlier if, in some lower dimensional projection, it is present in a local region of abnormally low density. More specifically, the present invention defines outliers in the data set by examining at those projections of the data having an abnormally low density. By defining clusters which are specific to particular projections of the data, it is possible to design more effective techniques for finding clusters.

Knorr discloses algorithms and applications for finding outliers using a distance based approach, and teaches that outliers are found by answering a nearest neighbor or range query with a specified radius for each object. If more than a specified number of neighbors are found within the range, or in the neighborhood of the object, it is declared a non-outlier. The object is declared an outlier if the number of neighbors found in the range is less than or equal to the specified number.

Knorr suffers from the inherent disadvantage of treating the data in a uniform way even though different localities of the data may contain clusters of varying density. When finding the outliers based on the density of their local neighborhoods and defining distances in full dimensional space, all pairs of points are almost equidistant and it becomes difficult to use these measures effectively.

Knorr focuses on finding outliers in multidimensional data sets, for example, "k-dimensional data sets with large values of k (e.g. $k \geq 5$)," (Abstract). However, Knorr does not focus on the high dimensional aspect of outlier detection, involving dimensions of 100 or 200, as in the present

invention. Therefore, Knorr uses methods which are more applicable for low dimensional problems, such as relatively straightforward proximity measures of which the complexity increases exponentially with dimensionality. Thus, for relatively smaller dimensions of 8 to 10, the technique of Knorr is computationally intensive. For higher dimensions, the technique is likely to be infeasible from a computational standpoint.

Regarding independent claims 1, 11 and 21, Knorr fails to disclose a technique for determining sets of dimensions and ranges in the data set which are sparse in density. The Examiner refers to the Abstract and paragraphs one and two of section 3.1 of Knorr in rejecting this element of independent claims 1, 11 and 21. However, a defined radius range query performed for each object does not provide the support necessary for an anticipation rejection since it differs significantly from a determination of one or more sets of dimensions and corresponding ranges which are sparse in density. As described above, while Knorr simply determines whether an object is an outlier by the number of neighbors found within a specified range, independent claims 1, 11 and 21 of the present invention recite the determination dimensions and corresponding ranges (e.g., projections) in the data set which are sparse in density (e.g., having an abnormally low density of objects).

In an Advisory Action sent on July 28, 2004, the Examiner contends that Knorr determines sets of dimensions and ranges during the process of determining distances between objects, since outliers are determined based on a range query and sets of dimensions corresponding to the ranges are determined. However, the present invention determines a set of dimensions which are sparse in density. The determined set of dimensions have corresponding ranges. Thus, the present invention does not begin with a range query, and Knorr fails to disclose the determination of a set of dimensions. This determination of a set of dimensions with corresponding ranges which are sparse in density provides for the identification of low density projections from the data.

Further, Knorr fails to disclose the identification of data points in the sets of dimensions and ranges as outliers. In response to Appellants previous arguments, the Examiner states that Knorr clearly teaches the identification of data in sets of dimensions and ranges. However, the Examiner fails to realize that since Knorr fails to disclose the determination of sets of dimensions and

corresponding ranges in the data set which are sparse in density, it also fails to disclose the identification of data points in these sets of dimensions and ranges as outliers.

Appellants assert that dependent claims 2, 3, 6-9, 12, 13, 16-19, 22, 23 and 26-29 are patentable for at least the reasons that independent claims 1, 11 and 21, from which they depend, are patentable. Further, dependent claims 2, 3, 6-9, 12, 13, 16-19, 22, 23 and 26-29 recite patentable subject matter in their own right. More specifically, dependent claims 6, 16 and 26 recite the determination of a set of dimensions using an algorithm which uses the processes of solution recombination, selection and mutation over a population of multiple solutions. In rejecting these claims, the Examiner refers to a section of Knorr describing index-based algorithms and analyses of multidimensional indexing schemes. However, Knorr fails to disclose an algorithm that uses the processes of solution recombination, selection, and mutation over a population of multiple solutions to determine a set of dimensions sparse in density. Dependent claims 7-9, 17-19 and 27-29 depend directly from claims 6, 16 and 26, respectively.

Accordingly, withdrawal of the rejection to claims 1-3, 6-9, 11-13, 16-19, 21-23 and 26-29 under 35 U.S.C. §102(a) is therefore respectfully requested.

(II) With regard to the rejection of claims 10, 20 and 30 under 35 U.S.C. §102(a) as being anticipated by Sheikholeslami, Appellants assert that Sheikholeslami fails to teach or suggest each and every element recited in claims 10, 20 and 30. The techniques of claims 10, 20 and 30 recite the detection of one or more outliers in a data set. One or more patterns in the data set are identified and mined which have abnormally low presence not due to randomness, and one or more records having the patterns present in them are identified as outliers.

Sheikholeslami discloses a wavelet-based clustering approach for spatial data in very large databases. More specifically, Sheikholeslami describes the discarding of noise objects during the mining process in stating that it is "insensitive to noise" (Abstract), thereby teaching away from independent claims 10, 20 and 30 of the present invention. Sheikholeslami finds and eliminates outliers as an aside-product of the clustering algorithms and does not use techniques focused on finding these deviations. Therefore, while Sheikholeslami describes a clustering algorithm that is able to identify clusters irrespective of their shapes or relative positions, it fails to disclose the

identification of patterns in the data set which have abnormally low presence not due to randomness in the detection of outliers.

Further, since Sheikholeslami fails to disclose the identification of such patterns having abnormally low presence, it also fails to disclose the identification of records as outliers that have the patterns present.

Accordingly, withdrawal of the rejection to claims 10, 20 and 30 under 35 U.S.C. §102(a) is therefore respectfully requested.

For at least the reasons given above, Appellants respectfully request withdrawal of the §102(a) rejection of claims 1-3, 6-13, 16- 23 and 26-30. Appellants believe that claims 1-3, 6-13, 16-23 and 26-30 are not anticipated by Knorr or Sheikholeslami. As such, the application is asserted to be in condition for allowance, and favorable action is respectfully solicited.

Respectfully submitted,

Date: August 10, 2004

Robert W. Griffith
Attorney for Applicant(s)
Reg. No. 48,956
Ryan, Mason & Lewis, LLP
90 Forest Avenue
Locust Valley, NY 11560
(516) 759-4547

## APPENDIX

1. A method of detecting one or more outliers in a data set, comprising the steps of:

determining one or more sets of dimensions and corresponding ranges in the data set which are sparse in density; and

determining one or more data points in the data set which contain these sets of dimensions and corresponding ranges, the one or more data points being identified as the one or more outliers in the data set.

2. The method of claim 1, wherein a range is defined as a set of contiguous values on a given dimension.

3. The method of claim 1, wherein the sets of dimensions and corresponding ranges in which the data is sparse in density is quantified by a sparsity coefficient measure.

6. The method of claim 1, wherein a set of dimensions is determined using an algorithm which uses the processes of solution recombination, selection and mutation over a population of multiple solutions.

7. The method of claim 6, wherein the process of solution recombination comprises combining characteristics of two solutions in order to create two new solutions.

8. The method of claim 6, wherein the process of mutation comprises changing a particular characteristic of a solution in order to result in a new solution.

9. The method of claim 6, wherein the process of selection comprises biasing the population in order to favor solutions which are more optimum.

10. A method of detecting one or more outliers in a data set, comprising the steps of:

identifying and mining one or more patterns in the data set which have abnormally low presence not due to randomness; and

identifying one or more records which have the one or more patterns present in them as the one or more outliers.

11. Apparatus for detecting one or more outliers in a data set, comprising:

at least one processor operative to: (i) determine one or more sets of dimensions and corresponding ranges in the data set which are sparse in density; and (ii) determine one or more data points in the data set which contain these sets of dimensions and corresponding ranges, the one or more data points being identified as the one or more outliers in the data set.

12. The apparatus of claim 11, wherein a range is defined as a set of contiguous values on a given dimension.

13. The apparatus of claim 11, wherein the sets of dimensions and corresponding ranges in which the data is sparse in density is quantified by a sparsity coefficient measure.

16. The apparatus of claim 11, wherein a set of dimensions is determined using an algorithm which uses the processes of solution recombination, selection and mutation over a population of multiple solutions.

17. The apparatus of claim 16, wherein the process of solution recombination comprises combining characteristics of two solutions in order to create two new solutions.

18. The apparatus of claim 16, wherein the process of mutation comprises changing a particular characteristic of a solution in order to result in a new solution.

19. The apparatus of claim 16, wherein the process of selection comprises biasing the population in order to favor solutions which are more optimum.

20. Apparatus for detecting one or more outliers in a data set, comprising:

at least one processor operative to: (i) identify and mine one or more patterns in the data set which have abnormally low presence not due to randomness; and (ii) identify one or more records which have the one or more patterns present in them as the one or more outliers.

21. An article of manufacture for detecting one or more outliers in a data set, comprising a machine readable medium containing one or more programs which when executed implement the steps of:

determining one or more sets of dimensions and corresponding ranges in the data set which are sparse in density; and

determining one or more data points in the data set which contain these sets of dimensions and corresponding ranges, the one or more data points being identified as the one or more outliers in the data set.

22. The article of claim 21, wherein a range is defined as a set of contiguous values on a given dimension.

23. The article of claim 21, wherein the sets of dimensions and corresponding ranges in which the data is sparse in density is quantified by a sparsity coefficient measure.

26. The article of claim 21, wherein a set of dimensions is determined using an algorithm which uses the processes of solution recombination, selection and mutation over a population of multiple solutions.

27. The article of claim 26, wherein the process of solution recombination comprises combining characteristics of two solutions in order to create two new solutions.

28. The article of claim 26, wherein the process of mutation comprises changing a particular characteristic of a solution in order to result in a new solution.

29. The article of claim 26, wherein the process of selection comprises biasing the population in order to favor solutions which are more optimum.

30. An article of manufacture for detecting one or more outliers in a data set, comprising a machine readable medium containing one or more programs which when executed implement the steps of:

identifying and mining one or more patterns in the data set which have abnormally low presence not due to randomness; and

identifying one or more records which have the one or more patterns present in them as the one or more outliers.